Pierre Haessig & Marie-Anne Lefèbvre – octobre 2018

Exploration de données temporelles

Cette séance de 2×1h30 a pour objectifs de vous entraîner à une activité rarement enseignée : explorer un jeu de données[1]*. Cette activité se situe* en amont *des traitements algorithmiques (régression, prévision, identification de modèle…) qui eux sont l'objet de divers cours (statistiques, machine learning…). L'objectif est de vous donner plus d'aisance pour démarrer vos prochains travaux sur des données (étude industrielle, stage…).*

À l'issue de cette séance, vous aurez un script permettant de traiter les données qui servent de support à ce cours (questions pratiques en orange), mais aussi des notes sur quelques concepts importants en science des données (questions théoriques et mots en bleu).

1 Chargement & prétraitements des données

Fichier : archive météo Rennes 2015

Reconnaître et charger différents types de fichiers (CSV, largeur fixe…)

Lister et classer en grandes familles les formats de fichiers de données tabulaires ?

Pour traiter des séries de données, il faut les mettre dans un type de *conteneur* adapté. Pour les traitements de données en Matlab c'est `table`, et son dérivé `timetable`.

Lister les différences entre une `table` et un tableau classique de Matlab ('array', matrice…) ?

(Attention : faire la différence entre le type du conteneur et le type du contenu : « verre vs eau »)

Quel est le format du fichier d'archive météo (cf. ISD-Lite 2006 format.pdf) ?

Charger l'archive de météo à Rennes en 2015 dans une `table`.
Extraire la température (5^{e} colonne, à diviser par 10)

Outils Matlab possibles : l'Import Tool[2] et fonction `readtable`. *NB : les dates sont pour l'instant non traitées.*

Bonne pratique : toutes vos manipulations doivent être **répétables automatiquement** (e.g. après un redémarrage de Matlab). Ccl : toutes les opérations doivent *in fine* être dans un script ou une fonction.

Données manquantes ou aberrantes

Avant de faire des analyses compliquées, une **exploration préliminaire** permet de voir à «quelles données on a affaire »

Tracer la série de température à Rennes (sur l'année, mais sans se préoccuper de la date).
Calculer la température moyenne (et min et max sur l'année).

Vous devriez remarquer quelques problèmes :

Y a-t-il des données aberrantes et/ou manquantes ? Si oui, combien et que faire ?

Comment sont représentées les données manquantes[3] sous Matlab ?

Remplacer les valeurs aberrantes par des valeurs manquantes (on peut utiliser le logical indexing[4]) et refaire le tracé et calculer la moyenne avec une fonction adaptée.

Observez : lorsqu'une longue série de données contient quelques valeurs manquantes, ça ne se voit pas sur un tracé (ex : faire un zoom autour de 2200) !

Dates : parsing (décodage) et représentation

Chaque fichier texte contient des dates représentées d'une façon particulière : '2018-10-22' ou bien '22/10/2018' (idem pour les heures). Matlab doit décoder (« *parser* ») ces chaînes de caractères pour en extraire le sens et stocker en mémoire l'information dans son format interne `datetime`[5] qui permet de faire des calculs avec (ex : durée entre deux dates).

Une fois la date décodée, le conteneur `table` peut être converti en `timetable`[6] qui facilite le traitement des données horodatées. Vous pouvez en particulier :

Tracer la température à Rennes sur le mois de juin
Tracer un `boxplot` des températures mensuelles. *NB : pour le cas particulier de cette archive météo, où l'on dispose d'une colonne numérique « mois », on pourrait se dispenser du décodage des dates si l'on veut juste ce tracé.*

2 Mise en forme & synchronisation de données

Fichiers : log capteur de qualité de l'air salle 404 (et archive météo)

Empilement/désempilement

Charger le fichier log des capteurs de qualité de l'air (humidité, CO2, COV, température intérieure et extérieure) dans un `timetable` (e.g. décoder la date et l'heure qui sont dans des colonnes différentes puis les recombiner[7]).

Que peut-on dire de la structure (*wide* vs *narrow*[8]) de ces données par rapport à l'archive météo ? Quels sont les intérêts de chacun de ces formats ?

Désempiler (unstack) les données pour tracer l'évolution de la température intérieure.

Statistiques par groupes

Calculer la température moyenne, min et max pour chaque mois et les afficher. Il faut pour cela utiliser les statistiques par groupes[9]

Rééchantillonnage et agrégation

Problématique : certaines analyses comme une régression statistique, nécessitent des données *synchronisées*. Un `timetable` peut être rééchantillonné (avec un pas plus rapide ou plus lent) avec `retime`.

Quels sont les pas d'échantillonnage des différentes données ? Les échantillonnages sont-ils réguliers ?

Tracer un nuage de point de la concentration en CO2 vs température intérieure.

Choix de la méthode de rééchantillonnage. Lorsqu'on veut sur-échantillonner, il faut choisir une méthode pour créer des valeurs qui n'existent pas : blocage d'ordre zero, interpolation linéaire… ? Lorsqu'on veut sous-échantillonner des données, on procède souvent à un *filtrage* ou une *agrégation statistique*. Il faut alors choisir la méthode d'agrégation (par ex. un moyennage ou une somme ?).

Tracer un nuage de point de la température extérieure (issue du capteur de la salle 404) vs la température à Rennes de l'archive météo.

3 Recap final

Fichiers : fiche de présence salle 404 (et log qualité de l'air, archive météo)

Charger la fiche de présence. Notez que ces données sont de nature différente que les précédentes (*catégorielle* vs *quantitative*). Quelle conséquence lors d'un sur-échantillonnage ?

Tracer un nuage de point de la température intérieure vs la concentration en CO2

Références

Fichiers

1) Archive de la station météo Rennes St-Jacques pour 2015. Fichier `meteo_rennes_2015.txt`

Source : NOAA Integrated Surface Database (ISD) https://www.ncdc.noaa.gov/isd

2) Log du capteur de qualité de l'air intérieur en salle 404. Fichier `log-20150309-171821.csv`

3) Fiche de présence en salle 404. Fichier `fiche_presence.csv`

NB : la salle 404 est la salle de réunion de l'équipe d'Automatique du campus de Rennes.

Liens

1. Exploratory data analysis. *Wikipedia* (2018).
2. Import Tool - Import data from file - MATLAB Documentation. Available at: https://fr.mathworks.com/help/matlab/ref/importtool-app.html. (Accessed: 5th October 2018)
3. Missing Data in MATLAB - MATLAB Documentation. Available at: https://fr.mathworks.com/help/matlab/data_analysis/missing-data-in-matlab.html. (Accessed: 22nd October 2018)
4. Logical indexing. *Steve on Image Processing*. https://blogs.mathworks.com/steve/2008/01/28/logical-indexing/
5. datetime - Arrays that represent points in time - MATLAB Documentation. Available at:

http://fr.mathworks.com/help/matlab/ref/datetime.html. (Accessed: 10th September 2018)

6. Timetables - MATLAB Documentation. Available at: https://fr.mathworks.com/help/matlab/timetables.html. (Accessed: 7th September 2018)
7. Combine Date and Time from Separate Variables - MATLAB & Simulink - MathWorks France. Available at: https://fr.mathworks.com/help/matlab/matlab_prog/combine-date-and-time-from-separate-variables.html. (Accessed: 10th September 2018)
8. Wide and narrow data. *Wikipedia* (2017).
9. Summary Statistics Grouped by Category - MATLAB Documentation. Available at: https://fr.mathworks.com/help/stats/summary-statistics-grouped-by-category.html. (Accessed: 22nd October 2018)